DOCUMENT RESUME

ED 157 057 CS 204 199

AUTHOR Butler, Roberta J.
TITLE Comparison of the University of Georgia Secondary English Education Evaluation Procedures with the Georgia Teaching Field Criterion-Referenced Test in Communicative Arts.
PUB DATE 78
NOTE 48p.; Research prepared at the University of Georgia

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *English Education; *Evaluation Methods; Higher Education; Language Arts; *Predictive Ability (Testing); Preservice Education; *Secondary School Teachers; Student Teaching; Teacher Attitudes; Teacher Certification; *Teacher Education; *Testing Programs; Test Validity

ABSTRACT

A study was undertaken to examine the traditional method of evaluating teachers, the multiple-evidence approach used by the University of Georgia's field-based preservice secondary English education program and by the Georgia Department of Education. The study sought to evaluate the instruments used by the university to measure prospective teachers' knowledge, attitudes, and performance, and to use these instruments in a comparative, evaluative study of the Georgia Teaching Field Criterion-Referenced Test in Communicative Arts. Subjects included 40 teacher candidates who had completed the university program, and 40 candidates who in addition had taken the state test as part of the test's trycut. The two groups scored alike on the university's test, establishing the groups' similarity and supporting the test's reliability. No significant relationship was found between the department's test and any other measure of performance, including professional or student ratings. The results indicate that none of the predictor measures is useful for positive prediction of classroom performance, and that these measures have little utility for prediction other than of performance on other measuring instruments. (DF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED157057

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Comparison of the University of Georgia Secondary English Education Evaluation Procedures with the Georgia Teaching Field Criterion-Referenced Test in Communicative Arts

Roberta J. Butler

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Roberta J. Butler

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

CS204 179

INTRODUCTION

The public's attitude toward education has changed little since Harry Broudy's description of the educational millieu of the 1960's. In 1972, he wrote:

> The public was told that teachers oppressed and murdered children (at least in spirit), and that the public school, like God, was dead. Throughout the decade the villain remained the same -- the mindless teacher allegedly produced by mindless education professors at mindless schools of education. (1972, p. 2)

Many people might argue that the description still applies. Parents, students, legislators, journalists, educators, or in other words, almost every sector of the American public is demanding that teacher preparation institutions and the agencies that certify graduates of those institutions be held responsible for the level of professional expertise possessed by the teachers they approve to teach. The accountability movement, the lack of available teaching positions, the low test scores publicized by the press, the law suits brought by illiterate yet diplomaed high school students, and many other forces have resulted in attempts to raise the standards for admitting people to the teaching profession. Thus, state departments and teacher preparation programs demand higher grade point averages, more courses, higher scores on teacher examinations, probationary rather than permanent certificates for first year teachers, and a variety of other requirements to raise professional standards. However, these steps to enforce more stringent standards for evaluation must be founded on sound educational, financial, research, philosophical, and moral principles in order to safeguard the interests of all the pressure groups already mentioned.

Those who prepare and certify teachers focus their evaluation either on the teacher or on that teacher's students. If they focus on the teacher, evaluators may try to assess 1) the teacher's knowledge and attitudes or 2) his/her teaching behavior or performance. When focusing on students, evaluators attempt to measure 3) growth in pupil knowledge and attitudes or 4) changes in pupil behavior. Some cautious evaluators attempt to collect all four types of information.

The Georgia Plan

The Georgia Department of Education and the University of Georgia's Language Education Department both choose to take a multiple evidence approach to the question of teacher evaluation, although still omitting assessment of pupil outcomes. The University of Georgia Language Education Department's evaluation includes measures of the teacher candidate's knowledge, attitudes, and performance; these measure will be described later. While the Georgia Department of Education's certification plans also involve a multiple evidence approach, evaluation will focus primarily on the teacher's knowledge and performance, and only indirectly on attitudes. The proposed changes are designed "to separate, in part, the certification and preparation functions by introducing job-related certification measures that are external to institutional grades, courses, and approved programs" (Georgia Department of Education, 1977, p. 1). This evidence includes local school system assessment of student teaching performance and a score on a criterion-referenced test in the prospective teacher's particular teaching field. This separation of preparation and certification along with the knowledge test constitutes a unique approach untried by any other state, according to a recent Phi Delta Kappan article (Villeme, 1977, p. 429).

In this Georgia plan, the prospective teacher must receive a degree from an approved teacher preparation program which includes student teaching in a local school system where local officials observe and rate the teacher candidate's performance. He/she must also take the criterion-referenced test developed by National Evaluation Systems, Inc. (NES) in contract with the Georgia Department of Education in the appropriate teaching field and must make a passing score in order to proceed any further in the certification process. If he/she passes, the prospective teacher receives a non-renewable certificate, valid for three years, and may teach if a job is available. During the first three years, local evaluation teams trained by the state department will observe and rate the new teacher's performance to determine if a regular renewable certificate should be issued. The conversion to a renewable certificate may occur at the end of any of the first three years, depending on how quickly the new teacher can demonstrate competence in the classroom (Georgia Department of Education, pp. 1-13). It has taken much time, work, and money to formulate these plans. However, it may take much more if a prospective teacher fails the criterion-referenced test and takes the case to court where the test might be judged invalid.

Statement of the Problem

The courts are gradually describing the guidelines for a selection process. In his book on certification practices, Tractenberg (1973) interprets what the courts seem to be saying in their decisions about the selection process. In the hearings on New York City's Board of Examiners' teacher certification exams, Stephen Pollak, former Assistant Attorney General in charge of the Civil Rights Division, warned of the serious legal question that is raised when certification is based on a written examination.

> "If a board refuses to hire, retain, or promote a teacher because of his score on a test, the board should be able to show that the test is a reliable predictor of the capacity of those taking the test to perform on the job in that system. If the board cannot make this showing, its action, if challenged will not be sustained.... Further, where a test measures only a portion of the qualifications required for successful performance on the job--and that is really

> true with all tests that I know of--and where members of a minority group uniformly score lower on the test, the Equal Protection Clause would preclude a school board from acting solely on the basis of the test....I fear that few, if any, school boards have made... the studies necessary to insure that a test serves their legitimate needs without discrimination. These studies must be made and repeated as needs change, if tests are to be the servant of the boards, rather than their master....Unless used within proper and careful limits, a test adopted as a part will become the whole of a selection process in what I believe will be serious risks of violations of the Constitution." (Tractenberg, pp. 255-256)

Since a prospective English teacher must pass the criterion-referenced test in Communicative Arts in order to win the opportunity to be observed and evaluated on the basis of classroom performance, the above cautions certainly apply.

In order to conduct the sort of study Pollak suggests, one would need to look at prospective teachers' scores on the criterion-referenced test and compare those scores with performance in real Georgia classrooms. In addition, one might be interested in comparing these test scores with scores on other measures of these prospective teachers' knowledge and attitudes. The Language Education Department carefully built into its pre-service teacher preparation program for secondary English teachers a wide range of evaluation prccedures which, as mentioned before, measure knowledge, attitudes, and performance. The instruments include a knowledge test, attitude inventories, and performance ratings, plus the traditional grade point averages and National Teacher Examinations scores used by the College of Education. The component of the program around which most of the evaluation revolves is the two quarter, field-based student teaching experience. A brief discussion of this experience is necessary at this point, but a more detailed description of the entire program, the evaluation procedures, and the instruments appears in Carol Angelia Moore's 1976 University of Georgia dissertation (see also English Education, Summer 1976).

For two quarters the prospective teacher works in several of the public middle and high schools which participate in one of the field-center operations.

During this time, the teacher candidate (TC) fills a variety of roles until finally he/she takes over full responsibility for teaching at least four classes for a period of eight to ten weeks. Both public school English teachers and college supervisors observe and rate the teacher candidate; the teacher candidate's pupils also rate his/her performance. Thus, there accumulates a rather complete picture of the teacher candidate's "on the job" performance.

Since the Language Education Department does utilize measures of knowledge, attitudes, and performance to evaluate pre-service teachers, and since these teacher candidates experience a situation similar to that of first year English teachers in Georgia schools, a study comparing their criterion-referenced test scores with their scores on the usual instruments employed by the Language Education Department would seem to be of value to state department, college of education, and public school officials, especially in light of earlier court decisions,such as Chance v. Board of Examiners in New York (Study Commission on Undergraduate Education and the Education of Teachers, 1976, p. 122), rejecting written examinations for certification purposes. However, a January 16, 1978 Supreme Court decision (Vlaanderen, 1978, p. 6) upheld a 1977 three judge federal district court panel ruling (McDaniel, 1977, p. 186) which permits South Carolina's continued use of the NTE as a certification test. Despite this precedent which establishes content validity as a sufficient proof of a test's job-relatedness, an opportunity to look at the relationship between student teachers' scores on the Communicative Arts field test and their scores on instruments designed to evaluate them as part of a field-based program would seem one that should not be ignored. Therefore, there are two purposes to this study.

1. To further validate the instruments and evaluation methods used in the University of Georgia's Pre-service Secondary English Education program.

2. To use these instruments in a comparative, evaluative study of the Georgia Teaching Field Criterion-Referenced Test in Communicative Arts.

Background

In order to justify and clarify the direction of this study, it was necessary to review the literature in three related areas: research in teacher effectiveness/evaluation, psychometric principles underlying criterion-referenced test construction, and legal issues related to the use of tests for selection purposes. However, space limitations require that this article discuss only research in teacher effectiveness, and since hundreds of studies exist , reputable reviews of the existing research will be cited in order to describe the overall state of knowledge about this area of research. For more detailed information, one can consult the comprehensive reviews cited or the original reports of the studies.

As explained earlier, one may identify four dimensions in which assessments can be made for teacher evaluation purposes: teacher knowledge and attitudes, teacher performance or behavior, student knowledge and attitudes, and student behavior. Evaluation experts and researchers disagree as to which of the four dimensions constitutes the best decision base for evaluating teachers; indeed, faults can be found with each of the four.

The first dimension, teacher knowledge and attitudes, is included in what Mitzel calls presage variables in his article in the 1960 Encyclopedia of Educational Research (p. 1482). Assessments of presage variables frequently involve paper and pencil tests of knowledge and attitudes, a method requiring attention to such psychometric concerns as reliability, validity, etc.

Another widely used method of assessing presage variables is examination of transcripts, the faults of which seem obvious.

Assessment of teacher performance or behavior is probably the most widely accepted method of evaluation at this time, but numerous researchers criticize evaluators who single out process criteria, as Mitzel calls them (p. 1482). These criticisms stem from the numerous technical difficulties involved in such assessments. Popham strongly criticizes process criteria, claiming that "the distressing truth is that systems for assessment and evaluation of teacher performance must be erected on technical foundations which more closely approximate balsa wood than concrete" (1971, p. 12).

In spite of their testimonies in the Second Handbook of Research on Teaching favoring pupil outcomes, or product variables as Mitzel labels them (Mitzel, p. 1482), McNeil and Popham (1973) quickly point out in the same article the difficulties in relying on growth in pupil knowledge and attitudes as a decision base for teacher evaluation. They admit technical and philosophical reservations about accepting pupil change as the criterion for judging teacher effectiveness (p. 218). Medley, Soar, and Soar discuss the philosophical reservations in somewhat stronger terms, concluding that evaluations based on pupil outcomes, whether pupil knowledge or behavior, stand on shaky moral ground as such evaluations force the teacher to violate the student's right not to learn (1973, p. 8).

Despite these attacks on the adequacy of each of the four dimensions as decision bases, teacher education and evaluation has always relied on these four kinds of variables. As Medley, Soar, and Soar explain, all of teacher education is based on the assumption that influences within each dimension have an appreciable effect on all the other dimensions. Indeed, the rationale behind the existence of teacher education programs is the belief that training

(presage variables) increases the teacher's effectiveness and affects pupil outcomes (product variables) (p. 3). However, when one looks to research to provide evidence that relationships among the four dimensions do exist, the picture which emerges seems less than convincing.

The research reviews will be presented in terms of relationships among the four types of variables. Since pupil knowledge and attitudes and pupil performance are usually included as a single category, this approach results in three categories of research: presage and process, presage and product, and process and product.

Presage and Process

The label presage criteria is applied because it assumes predictive power on the part of the criteria, some of which might be teacher intelligence, personality, training, experience, and social skills (Soar, 1962, p. 12). As pointed out earlier, teacher preparation programs are founded on the assumption that the training experiences affect the methods the teacher employs in the classroom. As Dunkin and Biddle explain, "although this is a reasonable presumption, we usually have no direct evidence to back it up" (1974, p. 49). Training a teacher to use a particular method does not guarantee that he/she will actually use that method in the classroom.

In summary, Flanders and Simon in the fourth edition of the Encyclopedia of Educational Research describe the progress in presage-process research as both sweet and sour.

> The sour is the continued difficulty of relating teacher traits to performance variables. In no study has sufficient attention been given to the methodological problems of keeping the conditions constant in the performance setting so that correlations with possible traits are maximized. Under the best of conditions reported, the correlations are still too low to give promise of utility; for example, errors in selecting prospective teachers by a predictor test would be too high except in the most extreme and obvious cases. (1969, p. 1431)

The example cited seems acutely relevant to this study.

Presage-Product

The relationship between teacher characteristics and student achievement has been the subject of a great deal of research, but according to Heath and Nielson's analysis of research reviews, "reviewers of this body of research generally conclude that an educationally significant relationship simply has not been demonstrated" (1974, p. 477). In 1963, Getzels and Jackson concluded that "very little is known for certain about the nature and measurement of teacher personality, or about the relation between teacher personality and teaching effectiveness" (p. 574). Heath and Nielson report the conclusions of reviewers of presage-product research, and as late as 1970, seven years after the Getzels and Jackson review, the situation was not much improved as illustrated by Mood's summary that "'we can only make the not very useful observation that at the present moment we cannot make any sort of meaningful quantitative estimate of the effect of teachers on student achievement'" (Heath & Nielson, p. 478).

Process-Product

Studies of this type abound and have warranted a number of thorough reviews. These studies measure, usually by direct observation, the "process" used by the teacher and measure, usually by mean gains on achievement tests, the "product" or student outcomes and attempt to correlate the two. Unfortunately, as the reviewers are quick to point out, the increase in process-product studies has not led to an increase in generalizeable conclusions about relationships between teacher behaviors and student achievement.

After reviewing 50 such studies, Rosenshine and Furst (1971) squeezed out only 11 variables which could boast empirical support. They conclude as follows:

> Of all the variables which have been investigated in process-product studies to date, five variables have strong support from correlational studies and six variables have less support but appear to deserve future study. The five variables which yielded the strongest relationships with measures of student achievement are: clarity, variability, enthusiasm, task orientation and/or business like behavior, and student opportunity to learn. The six less strong variables are: use of student ideas and/or teacher indirectness, use of criticism, use of structuring comments, use of multiple levels of discourse, probing, and perceived difficulty of the course. The relationships are positive for ten of the variables and negative for use of criticism. (p. 54)

Thus the process-product research does not yield overwhelming evidence upon which to base teacher evaluation decisions; the studies yield only eleven competencies, and of these only five seem to be valid according to Rosenshine and Furst.

A second review makes the situation appear even more futile. Heath and Nielson reanalyzed 42 of the studies cited by Rosenshine and Furst and left the validity of even the strongest five variables in question. Heath and Nielson argue that their results deny the existence of an empirical basis for performance-based teacher education; their conclusions speak directly to teacher educators and certifiers.

> First, the research literature on the relation between teacher behavior and student achievement does not offer an empirical basis for the prescription of teacher training objectives.
>
> Second, this literature fails to provide such a basis, not because of minor flaws in the statistical analysis, but because of fundamentally weak research designs.
>
> Last, given the well-documented, strong association between student achievement and variables such as socioeconomic status and ethnic status, the effects of techniques of teaching on achievement (as these variables are defined in the PBTE research) are likely to be inherently trivial (p. 481)

Summary

In light of the discouraging results obtained from process-product studies, N. L. Gage (1974) points out that despite its popularity, process-product research is not the only research worth conducting. He argues that achievement constitutes only one end of a causal chain or network and

is not the only justification for teaching procedures or teacher education. Using the analogy of the physician whose competencies are not validated solely on the basis of their correlation with patient longevity or recovery, Gage claims that:

> Many desirable factors in teaching or teacher education may be too remote from that final link in the chain to have a discernible effect upon it. They are desirable, nonetheless, if they can be shown to have good effects on the links closer to the treatment, or independent variable, that we are evaluating. (p. 181)

Many other educators agree with Gage as to the inappropriateness of limiting teacher evaluation to a narrow field of criteria. Realizing the infirmities of the research base for teacher effectiveness, many researchers suggest redirecting research efforts based on a more realistic conceptualization of teaching, and there seems to be a fair amount of agreement on this direction.

Bruce Joyce reiterates the view that teaching "is a multidimensional act that fits in a multidimensional context to produce multidimensional environments which are interacted with by multidimensional learners" (1974, p. 62) and admonishes those who expect univariate truth from research to result in mechanistically termed formulas prescribing a certain teaching skill to increase a certain kind of learning. This view is shared by at least two other reviewers of teacher effectiveness research. Harootunian laments the likelihood of payoff if studies continue in the same direction and offers instead a new perspective which suggests looking for "combinations of learner aptitudes and educational treatments which will produce learner effects" (1975, p. 128). Instead of looking for the set of competencies which describes the effective teacher, Harootunian says researchers should ask "what characterizes the effective teacher for students with characteristic A as opposed to one for students with characteristic B?" (p. 131). Similarly, Medley complains that "past research has developed so little knowledge of the nature of effective teaching that I have had to propose a model for teacher education which does not assume the existence of any such knowledge" (1975, p. 173). He charges

that we should cease regarding teacher effectiveness as a unidimensional trait and teacher competence as a static concept. Future research should provide descriptive information which will help teachers choose a particular competency in a particular situation. "It will accomplish its purpose by asking, not which procedure effective teachers use most frequently, as it does now; but instead it will ask under what conditions an effective teacher uses this procedure, for what purposes, with what kind of pupils, and so on" (Medley, 1973, p. 44).

Until research does follow this new direction and provide a firmer knowledge base, at least one group of educators warns against using "specific levels of teacher performance as 'go' or 'no go' criteria for entry into or promotion within the profession" (Ward, Morine, Berliner, 1974, p. 124). This warning and the suggestions of Joyce, Harootunian, and Medley echo Gage's summation that "proximate criteria of all sorts are justified" (p. 182) in determining how to assess teacher effectiveness. The opinions of these experts in this field of research offer support for the multiple evidence approach to teacher evaluation taken by the Language Education Department at the University of Georgia and by the Georgia Department of Education.

Research Design

As already explained, traditionally there are four categories of evidence upon which to base evaluations of teachers. The first two categories are (A1) teacher knowledge and attitudes and (A2) teacher performance or behavior. The last two categories are (B1) pupil knowledge and attitudes and (B2) pupil behavior. The review of the literature related to teacher effectiveness/evaluation research concludes that one should not insist that all measures of teacher effectiveness be based on product variables or pupil outcomes. Moreover, several of the researchers cited agreed that all types of evidence should be utilized.

In accordance with the views expressed by these researchers, the University of Georgia Language Education Department makes use of a variety of measures of teacher knowledge, attitude, and performance in evaluating prospective teachers and its own Pre-service Secondary English Education program. The kinds of measures used include a knowledge test, attitude inventories, performance ratings, pupil ratings, the NTE, and GPA's. The initial evaluation, which also included a follow-up study, was conducted by Moore in her previously cited 1976 study. The present study, since it also dealt with students in this same program, utilized the same instruments and thus provided data which was used as a basis for comparing the Moore group and the present group of TC's for the purpose of further confirming the validity and utility of these instruments for beginning teacher evaluation and, indirectly, for program evaluation. Once this step was accomplished, these measures were used to examine the Georgia Teaching Field Criterion-Referenced Test in Communicative Arts (CRT-CA). Since these measures supply evidence which falls into two of the four categories, it was possible to compare the CRT-CA, a measure of teacher knowledge, with other measures of teacher knowledge as well as with measures of teacher attitudes and teacher performance.

The following categorization identifies the measures and the dimensions into which they fit.

A1) Teacher Knowledge: Departmental English Education Test (DET), Georgia Teaching Field Criterion-Referenced Test in Communicative Arts (CRT-CA), National Teacher Examinations (NTE), and major and overall GPA.

Teacher Attitude: Minnesota Teacher Attitude Inventory (MTAI) and Fundatmental Interpersonal Relations Orientation -- Behavior (FIRO-B)

A2) Teacher Performance or Behavior: Teacher Candidate Progress Report Form (TCPR) and Inventory of Student Perception of Instruction. (ISPI)

B1) Pupil Knowledge and Attitude: None

B2) Pupil Performance or Behavior: None

All of the data for these measures were recorded for statistical treatment. Although the statistical treatment was not conducted in an attempt to infer absolute causal relationships, it did allow for statistical analysis of relationships and differences between groups and among measures.

The population in the Moore study consisted of 40 students who completed the two quarter field-based student teaching program from January 1975 to June 1976. These 40 students made up three field center groups. The population for the present study consists of 40 students who completed the program from September 1976 to June 1977. These students made up four field center groups.

Assignments to field centers depend on time and scheduling; thus the groups are basically self-selected. For this reason, no attempt is made by the department to match them for experimental purposes. In view of the previous information, it can be argued that the 1975-76 group from the Moore study and the 1976-77 group are quite similar since all of the students in both groups were prospective English teachers at the same stage in their academic and professional preparation and were submitted to the same screening procedures before being admitted to the program.

Questions

Based on the notions of teacher evaluation, criterion-referenced test development, and legal and psychometric validity, several related questions were asked in this study.

1. How do the 1976-77 TC's correspond to the 1975-76 TC's in terms of final scores?
2. Do certain teacher knowledge, attitude, and performance scores cluster, thus enabling a reduction of scores into principal component or factor scores?
3. What is the relationship between each of the measures and the professional ratings of performance?
4. Is there some combination of measures that will best predict performance as measured by the professional ratings?
5. What is the relationship between each of the measures and the estimated score on the CRT-CA?
6. Is there some combination of measures that will best predict performance on the CRT-CA?

The specific instruments used to answer these questions are described in the author's 1978 University of Georgia dissertation and in the Moore study (1976). The development of the CRT-CA is discussed by Nassif & Dewitt of NES in The Georgia Teaching Field Criterion-Referenced Test Project: Final Report, where they conclude that "these tests easily meet high quality measurement standards" (1977, Appendix F). The CRT-CA focuses on content knowledge only. The test consists of nine domains and 250 multiple-choice items. The domains include logical reasoning, research, composition, language, communications, communications careers, literature, reading, and oral communications.

Here it seems useful to summarize the research design in the following table which matches the questions, the instruments for collecting data, and the appropriate analysis. Only the post-test scores from these instruments were used, and two of these scores functioned as both predictor and criterion measures for certain analyses.

Table 1

Summary of the Research Design

Question Number		Instrument	Analysis
1	Attitude	a. MTAI FIRO-B expressed inclusion expressed affection expressed control wanted inclusion wanted affection wanted control	26 t tests for significance of differences between means
	Knowledge	b. NTE total common optional GPA overall major DET	
	Performance	c. TCPR preteaching preparation initiating behavior mode of presentation learning activities classroom management professional attitude overall performance	
	Performance	d. ISPI instructional objectives human relations instructional resources motivation meaningfulness of learning measurement and evaluation	
2		All of the above	principal component analysis on each of the four groups of instruments

Table 1

Continued

Question Number	Instrument	Analysis
3	Component or factor scores on all of the above plus the CRT-CA estimated total score	intercorrelations
4	All of the above	stepwise multiple regression
5	All of the above	intercorrelations
6	All of the above	stepwise multiple regressions

Insert Table 1 about here

The t tests were used to compare differences in mean scores. The principal component analyses were conducted with the expectation that the number of measures could be reduced to a more workable number of component or factor scores.

The four correlational analyses involve the CRT-CA, and, there are two sets of CRT-CA scores corresponding to two separate groups of TC's who took two separate forms of the field test. According to Angoff, there is a method to predict scores on one form of the test from the scores on another form of the test if the two forms are equally reliable (Angoff, 1971, p. 569). The formula used is a conversion equation, which made it possible to obtain for each student a score for each half of the CRT-CA; these two scores were added together to provide an estimated total CRT-CA score for each of the 40 TC's. Before this method could be used, three comparisons had to be made between the two test forms. First, there was a rational content analysis of the two tests, a comparison of the average difficulty of the two forms, and a comparison of the reliability as reflected by the standard error of measurement for the two forms.

Three judges compared the two test forms to determine the similarity of content. NES had counterbalanced the items in each booklet to insure an equitable distribution of content matter, resulting in two test forms both of which contained items from eight of the nine domains. Only one domain was confined to one test form, and only one item from that domain appeared in the test booklet. Thus the judges concluded that the overall distribution of content was equivalent. The average

difficulty of the two forms was computed, and the difference in these two values was only .02, justifying the conclusion that the two forms were equivalent with respect to difficulty. Each test form's standard error of measurement was obtained, and the two values differed by only .70, permitting the conclusion that the two forms were equally reliable.*

These comparisons justified the use of the conversion equation; thus, estimated total scores were used throughout. The total score refers to the student's performance only on the selected items, not to all the items which appeared in the field test booklet. Items later rejected by NES as unsatisfactory were omitted from the computation of the student's score.

The Procedures

The instruments described above were administered to the TC's at various times during the program. The procedures for gathering and recording the data were as follows. The final TCPR was written and filed during the last week of student teaching for use in a final conference with the TC. These progress reports were based on at least 10 and sometimes as many as 35 classroom observations on each TC. During the last week of the college quarter, the ISPI was administered by the supervising teacher in two of the classes for which the TC had primary responsibility. The ISPI was machine scored and recorded after the TC had left the school.

On one of the last days of the second quarter of the student teaching experience, each field center group met on the University campus. This meeting lasted for approximately three hours during which time the TC's completed the following instruments: MTAI post-test, the DET, and FIRO-B post-test. The order in which the tests were administered was changed with

*Georgia Department of Education and National Evaluation Systems, Inc. currently have contracts in force which prevent the report of exact statistics.

each group. Finally, these instruments were hand scored and recorded for the appropriate statistical analysis. The NTE scores and the GPA's were obtained from the College of Education's official records. The field test CRT-CA scores were obtained from NES in cooperation with the Georgia Department of Education and were used to compute the estimated total scores as described previously.

Limitations

The limitations of the study include subjects, instruments, and general limitations of the design. First, the subjects were limited, for certain analyses, to those 80 students enrolled in the English Education program at the University of Georgia from January 1975 through June 1977 and, for other analyses, to those 40 students enrolled from September 1976 through June 1977. Because of the limited number enrolled in this program, the results do not necessarily generalize to other groups. Secondly, in a few cases, either the reliability or the validity of the instruments needs further verification. Additionally, the fact that the scores on the CRT-CA are scores obtained on the field test version may be a limitation. Finally, in the design, cause and effect relationships are not identifiable. Moreover, the fact that the students took pre-tests of some of the instruments might influence their post-test scores, thus possibly restricting the generalizeability of the results.

Results of the Study

This article reports the results of the statistical analyses conducted in an attempt to answer the research questions which dealt first with the comparison of the 1976-77 group of TC's with the 1975-76 group, next with the reduction of the number of variables, and finally with the correlational analyses of the relationship between and among measures of knowledge, attitude, and performance, and the CRT-CA score.*

*Due to limitations on the length of articles, statistical assumptions and complete data tables have been omitted but are available from the author.

Answers to the Research Questions

1. How do the 1976-77 TC's correspond to the 1975-76 TC's in terms of final scores?

For the purposes of this study, to claim that there were no differences between the two groups when differences did in fact exist (a Type II error) would constitute a greater hazard than the reverse (a Type I error). Therefore, in spite of the large number (26) of tests, the statistical analyses used to answer this question were *t* tests of significance for a difference between two means. These tests were conducted on the mean scores of the 1975-76 group and the 1976-77 group on the measures designated in Table 1. Two-tailed tests of significance were used, and the level of significance was set at .10.

The *t* tests indicate that the 1976-77 group and the 1975-76 group differ significantly on 8 out of 26 measures. Due to the type of significance test and levels used, there is little likelihood that the groups differ significantly on any other measure. Due to the fact that such a large number of *t* tests were conducted, there is a greater likelihood that the claim of significance for these eight differences may be erroneous, and these differences may not exist (a Type I error). In view of this possibility and of logical considerations already discussed concerning the similarities of the two groups, the only justifiable statement about the comparison of the two groups seems to be that statistically significant differences may exist on some of the variables, but in reality the two groups are probably very similar. In summary, the two groups seem similar enough to permit conclusions about the validity and utility of the package of instruments which the Language Education Department uses in evaluating prospective teachers.

2. Do certain teacher knowledge, attitude, and performance scores cluster, thus enabling a reduction of scores into principal component or factor scores?

In order to conduct the correlational analyses called for by the last four research questions, it was necessary to somehow reduce the number of variables from the original 26 to a smaller number of principal component or factor scores. Such reduction has been endorsed by a number of researchers familiar with teacher evaluation/effectiveness studies; one of these is Soar, who writes, "The need for analysis and synthesis to reduce the number of traits believed to be important is evident" (1962, p. 17). Based on inter-correlations among the variables, factor analysis attempts to identify clusters of common variance in order to determine how many factors are at work in a particular setting and the nature of these factors. This information enables the researcher to explain the variance in a situation by using fewer factors (Veal, Pikaart, and Berryman, 1967, pp. 16-22). For these reasons, four principal component analyses (PCA) were conducted: one for the seven attitude measures, one for the six knowledge measures, one for the seven performance measures on the TCPR ratings by the supervisors, and one on the six performance measures on the ISPI ratings by students. These analyses resulted in three, two, one, and one factors respectively.

Because the principal component analysis procedure requires approximately 10 times as many subjects as variables, scores for both groups of TC's were included, resulting in a total of 80 subjects, a number large enough to satisfy the requirement. The statistically significant differences cited in the previous section were considered, but the overall similarity of the groups, as discussed in the preceding section, seemed adequate enough to override the statistical cautions and favor the logical considerations.

Each of the four PCA's generated varimax rotated factor matrices which show factor loadings for each variable. Variables with factor loadings of less than .40 were omitted since such loadings are difficult to interpret as

they indicate that a very small percentage (less than 16%) of the variance of the instrument has been accounted for in the matrix. Loadings of the variables on the factors are used to identify and to describe the factors.

The PCA of the MTAI and the six FIRO-B scales extracted three attitude factors identified as Social Orientation, Acceptance of Responsibility and Leadership, and Beliefs about Students and Teaching. The knowledge PCA of the DET, NTE total, common, and optional scores, and the major and overall GPA's extracted only two factors: Tested Knowledge and Academic Success. Using the seven categories of the TCPR, the PCA extracted only one factor, identified as Professional Ratings. Finally, the PCA of the six ISPI sections also resulted in a single factor, Student Ratings.

In summary, certain knowledge, attitude, and performance scores do cluster, and the four PCA's extracted seven factors, three related to attitude, two to knowledge, and two to performance, a considerable reduction from the original 26. The procedure also generated factor scores for each of the 80 TC's on each of the seven factors. The factor scores of the 40 TC's in the 1976-77 group were used in the correlational analyses discussed next.

The correlational analyses were conducted in an attempt to discover the relationship between each of the measures and two separate criterion measures, one a performance rating and one a knowledge test score, both of which serve as predictors for each other.

When discussing results of regression analyses, the reliability of the instruments used as criterion measures is of utmost importance. The TCPR summarizes information contained on some 15-35 TC Observation Forms completed by at least three to five different members of the supervisory team. This TC Observation Form is a hybrid instrument in that it incorporates a rating scale into a form used by live observers to record classroom occurrences for the

purpose of discussing with a TC his/her performance in a particular class. The final TCPR form is used in the TC's final conference with supervisory team members to discuss overall strengths and weaknesses in his/her performance. Evidence from composition research as to numbers of raters (Veal & Biesbrock, 1969, p. 7) confirms that the number of persons observing the TC's and completing the TC Observation Form insures satisfactory rater reliability. However, due to the nature of the TCPR as a summary of a process, none of the usual methods for obtaining a reliability coefficient seems appropriate. While the omission of the reliability coefficient poses an inconvenience for the researcher, the instrument remains in use because of its utility in helping the TC's improve their performance, and, after all, it is this improvement which governs the choice of instruments, not a researcher's convenience. The reliability coefficients for the two forms of the CRT-CA were computed, and both are higher than .85. As explained earlier,contracts currently in force prevent the report of the exact statistics.

The result is that for the criterion Professional Ratings, the predictors are the six other factors and the CRT-CA. For the criterion CRT-CA, the predictors are the seven factors identified by the principal component analyses. The PCA's did yield factor scores for all 80 TC's; however, only the 40 1976-77 TC's took the CRT-CA. Therefore, only these 40 subjects could be used in the stepwise multiple regression procedures. Since there are seven predictor measures involved in each of the two multiple regressions, the number of subjects is somewhat smaller than would be ideal. However, logic would suggest that useful information could still be obtained from such analyses, if the statistical cautions are kept in mind when intepreting results. In testing the statistics from the regressions, two-tailed tests of significance were used throughout.

3. What is the relationship between each of the measures and the professional ratings of performance?

The stepwise multiple regression generated an intercorrelation matrix which provides the answer to this research question.

Insert Table 2 about here

As the table shows, there is only one significant correlation coefficient. The factor Tested Knowledge correlates significantly with the Professional Ratings of performance. None of the other coefficients are significant at the .01 level or at the .05 level. It is important to note that the significant correlation coefficient is -.58, indicating a negative relationship between the two variables. This negative coefficient indicates that persons who have a high Tested Knowledge factor score also have low Professional Rating factor scores and vice versa.

4. Is there some combination of measures that will best predict performance as measured by the professional ratings?

A stepwise multiple regression was conducted in an attempt to answer this question. With the Professional Ratings as the criterion, the seven variables used as predictors were the other six factors and the CRT-CA score. The following table gives the results.

Insert Table 3 about here

The only significant F, the statistic which tests the significance of the added variable, is the extremely high one associated with Tested Knowledge. Further, by examining the R square and the R square change columns, one can

Table 2

Correlations Between Predictors and Professional Ratings

	Professional Ratings
Social Orientation	.24
Acceptance of Responsibility and Leadership	-.03
Beliefs about Students and Teaching	-.08
Tested Knowledge	-.58**
Academic Success	.05
Student Ratings	.04
CRT-CA	-.26

*significant at the .05 level
**significant at the .01 level

Table 3

Stepwise Multiple Regression Summary Table Showing Significant F Values

Step	Variable	F	Multiple R	R Square	R^2 Change
1	Tested Knowledge	15.537**	.578	.334	.334
2	Academic Success	2.391	.619	.383	.049
3	Social Orientation	.657	.630	.397	.014
4	Beliefs about Students and Teaching	.992	.646	.417	.021
5	Student Ratings	1.019	.662	.438	.021
6	Acceptance of Responsibility and Leadership	.400	.669	.447	.008
7	CRT-CA	.310	.674	.454	.007

*significant at the .05 level
**significant at the .01 level

see that little is added by including any more of the variables in the regression. Using the Fisher z transformation, a confidence interval for the correlation coefficient was set up. This interval shows the upper and lower limits of the probable population R: .95 CI $[-.76 \leq r \leq -.33]$. This negative relationship (-.578) indicates that TC's who achieve high scores on Tested Knowledge tend to receive lower ratings from supervisors; in contrast, students who receive high professional ratings tend to do less well on knowledge tests.

5. What is the relationship between each of the measures and the score on the CRT-CA?

In order to answer this question, one must again look at results from the intercorrelation matrix. With the CRT-CA as criterion, the seven factors serve as predictors.

Insert Table 4 about here

This table reveals that there are only two predictors which correlate significantly with achievement on the CRT-CA. Both of these predictors are knowledge factors; Tested Knowledge and Academic Success both have a positive relationship with the CRT-CA. These high correlations do not seem surprising since the CRT-CA is a test of subject matter knowledge, and one would expect that a person with a high GPA and high NTE scores would also achieve a high score on another knowledge test.

6. Is there some combination of measures that will best predict performance on the CRT-CA?

To answer this last research question, another stepwise multiple regression was conducted, this time with the CRT-CA as the criterion measure and the

Table 4

Correlations Between Predictors and CRT-CA

	CRT-CA
Social Orientation	-.29
Acceptance of Responsibility and Leadership	.01
Beliefs about Students and Teaching	.03
Tested Knowledge	.60**
Academic Success	.46**
Professional Ratings	-.26
Student Ratings	.21

*significant at the .05 level
**significant at the .01 level

seven factors as the predictors. The results appear in Table 5, which follows.

Insert Table 5 about here

This table shows that the two knowledge factors are the only variables with significant F values, and the R square and R change columns indicate that little is added by including other predictors. As pointed out earlier, the small number of individuals (40) involved in this multiple regression demands that statistical precautions be taken in interpreting this multiple R (.671). There are methods for estimating the population R which take into account the small sample size. Two of these formulas are easily computed and can be applied in this situation; both of these formulas estimate shrinkage in the multiple correlation coefficient thereby enabling the researcher to more accurately predict the population correlation coefficient. The first formula, devised by John Garland Claudy (1969, p. 69), computes the estimated cross-validated $\hat{Rho}$ and, in this situation, yields .529. The second method, derived by Uhl and Eisenberg (1970, p.488), involves computation of the estimated multiple correlation coefficient $\bar{R}$ after shrinkage and yields .472. Therefore, despite the small sample size, the multiple correlation coefficient still indicates a significant relationship in the population.

In short, the combination of these two knowledge variables, Tested Knowledge and Academic Success, provides the best predictor of scores on the CRT-CA. The relationship is positive, indicating that those TC's who have both high GPA's and high NTE scores tend to do better on the CRT-CA than those with lower GPA's and NTE scores.

Table 5

Stepwise Multiple Regression Summary Table Showing Significant F Values

Step	Variable	F	Multiple R	R Square	R^2 Change
1	Tested Knowledge	17.503**	.601	.361	.361
2	Academic Success	4.875*	.671	.450	.089
3	Social Orientation	.753	.681	.464	.014
4	Beliefs about Students and Teaching	.583	.689	.475	.011
5	Acceptance of Responsibility and Leadership	.638	.698	.487	.012
6	Professional Ratings	.291	.702	.493	.006
7	Student Ratings	.034	.702	.494	.001

Conclusions

Conclusions Related to the Purposes of the Study

As already stated, the purposes of this study were to further examine the measures used by the Language Education Department to evaluate teacher candidates, and having found these measures to be valid, to use them to evaluate the Georgia Teaching Field Criterion-Referenced Test in Communicative Arts (CRT-CA). All of the conclusions in this section relate to these purposes.

Results of the analyses reported reveal that, as expected, the 1975-76 group of TC's and the 1976-77 group of TC's appear to be similar when mean scores on knowledge, attitude, and performance measures are compared. This similarity seems a logical expectation considering the requirements for admission to the program and the fact that these students have completed the same preparation program involving similar experiences in similar settings. Had this comparison revealed a larger number of differences between the two groups, the reliability of instruments used might have been called into question. However, since the groups differ on only a small number of measures (8 of 26), the most reasonable conclusion seems to be that the measures are not only reliable, but sensitive enough to measure minor differences between groups.

Moreover, these instruments apparently measure what they purport to measure since the PCA's revealed a limited number of logical clusters of variables. Had a particular PCA extracted a large number of factors, one might question the validity of the instruments involved. Since, however, the seven attitude measures clustered into only three factors, the knowledge measures into only two factors, the seven TCPR categories into one factor, and the ISPI sections also into one factor, the PCA's indicate that the instruments are indeed valid.

In summary, the comparison of the two groups of TC's provides further evidence as to the reliability of the instruments used to measure knowledge, attitude and performance of the TC's. The clusters of measures which emerged in the PCA's provide further evidence of the validity of these measures. As a result of this further evidence, these instruments seem reliable and valid enough to justify their use in evaluating the CRT-CA.

The relationships among all the eight variables (seven factors plus the CRT-CA) are described by the intercorrelations computed along with the multiple regressions. Only three correlation coefficients are significant; there exists a positive relationship between the two knowledge factors, Tested Knowledge (.60) and Academic Success (.46), and the CRT-CA and a negative relationship (-.58) between Tested Knowledge and Professional Ratings of performance. The stepwise multiple regressions reiterated these relationships and indicated that a combination of Tested Knowledge and Academic Success scores provide the best prediction of scores on the CRT-CA and that no combination of scores provides prediction of Professional Ratings. The single best predictor of performance as measured by Professional Ratings is the Tested Knowledge factor whose negative correlation coefficient indicates that those TC's who have high Tested Knowledge scores received lower performance scores from supervisory team members than did those TC's who have lower Tested Knowledge scores. This evidence forces the conclution that none of the three attitude factors nor the Student Ratings function effectively as predictors of performance in the classroom as measured by Professional Ratings or of performance on the CRT-CA knowledge test.

Further, the CRT-CA does not predict performance as measured by Professional Ratings, nor does Professional Ratings or Student Ratings of performance predict success on the CRT-CA. There exists no significant relationship between the CRT-CA and any of these measures of attitude or performance. Tested Knowledge

and Academic Success together predict scores on the CRT-CA. Both these predictors are knowledge factors, and CRT-CA is a knowledge test. In short, GPA's combined with scores on one knowledge test predict scores on another knowledge test, a finding which arouses little surprise.

Conclusions Related to Teacher Effectiveness Research

The failure of this study to uncover more relationships among the eight variables and between the predictors and criterion measures merits attention but little surprise or disappointment when one considers the general picture which has emerged from previous research of Teacher Effectiveness/Evaluation. Studies attempting to relate presage variables, especially teacher characteristics such as knowledge and attitudes, to process and product variables have accomplished, according to Dunkin and Biddle, "strikingly little" (1974, p. 441). Other researchers concur with this assessment as indicated by the reviews already cited.

The negative relationship between NTE scores and the Professional Ratings of performance may seem startling at first, but previous studies have found negative or low positive correlations between the NTE Common Exam score and ratings by college supervisors and principals during student teaching (Quirk, et. al., 1973, p. 107). Moreover, other studies offer contradictory evidence as to the relationship of the NTE with process variables. Dunkin and Biddle (1974) cite findings from studies which report that "indirectness," a variable they conclude is weak and should be abandoned in future research, is associated with higher teacher scores on the NTE history and philosophy examination (p. 362). In contrast, they cite findings from Medley and Hill (1970) which report that "teacher criticism," a variable Dunkin and Biddle conclude is most promising for further research, is associated with higher scores on the NTE English A optional and the common literature examinations (p.368). This association between high NTE English

and literature examinations and criticism, one of the few variables which has been shown by research to have a relationship with pupil achievement (albeit a negative one), may help to make the negative association between NTE and performance seem less shocking. Certainly a TC who offers frequent criticism to students in the classroom would receive lower ratings from supervisors. In sum, the comparison of this study with previous studies suggests that the Professional Ratings are based either on an unreliable instrument or an instrument more sensitive than ones used in previous studies.

While results from previous research may allay some of the confusion generated by this particular finding from this study, studies using presage variables have refuted the logical expectation that one could use such instruments as the MTAI and NTE to predict teacher performance in the classroom. Dunkin and Biddle (1974) offer one explanation of the failure and a suggestion for development of future instruments.

> Why have we so far failed to find teacher properties that would predict classroom events? We suspect that the basic difficulty is that inventories for the measurement of teacher properties were developed independently of knowledge concerning actual classroom behavior....
>
> One alternative procedure might be to begin with observed classroom behaviors and to work backwards from them to the development of paper-and-pencil instruments. (pp. 441-442)

The positive relationship between the two knowledge factors and the CRT-CA, all presage variables, poses no challenge to logic as one would expect to be able to predict performance on a knowledge test from prior performance on knowledge tests and from grade point averages. Moreover, the CRT-CA's failure to predict performance as measured by Professional Ratings ($r = -.26$) falls prey to the above questions asked by Dunkin and Biddle of other presage-process research, and leads to a discussion of testing concerns.

Conclusions Related to Testing of Teachers

The CRT-CA seems to meet all the requirements of a carefully constructed criterion-referenced test. The account by Nassif & Dewitt (1977) of the development of the test provides evidence that NES followed accepted procedures to insure reliability, job relatedness, and content validity of the final 250 item version of the test.

Despite its careful development, the test does not possess demonstrated predictive validity. For this reason, perhaps the Georgia Department of Education should reconsider the CRT-CA's place in the certification procedure. As planned, the CRT-CA is a major hurdle in the certification process. Conceivably, a TC could complete an undergraduate program and receive outstanding performance ratings in student teaching, yet never be allowed the chance to be evaluated on the basis of performance in a classroom as a first year teacher if he/she does not achieve a high enough score on the CRT-CA. Therefore, the utility of the CRT-CA may be questioned for several reasons in light of the results of this study and recent court rulings.

Conclusions Related to Legal Concerns

This study included too few ethnic minority subjects to allow any specific conclusions as to any possible discriminatory impact resulting from the CRT-CA. If, when implemented, the test does prove to have a discriminatory impact, it may be challenged; there is at present no evidence to demonstrate its predictive validity. Nevertheless, the court may not deem such evidence necessary in light of the recent Supreme Court ruling upholding South Carolina's use of the NTE. The NTE did surmount one legal hurdle which in the future may force a comparison of the CRT-CA and the NTE. Often applied by the courts in such litigation, the "compelling state interest" or "business necessity" test requires the defendant to demonstrate both that the exam is job-related and that no less discriminatory means of selection exists (Tractenberg, 1973, p. 200).

The NTE has withstood an extensive content validation study through the jury judgement method and been upheld by the Supreme Court as a legally sound selection examination. Further, a number of colleges and schools of education require students to take the NTE; therefore, these scores could be readily available to certification officials. Since the NTE scores do correlate highly with the CRT-CA scores, a correlation which indicates concurrent validity, perhaps it would be cheaper and easier to reinstate the NTE as the certification exam if its discriminatory impact is no greater than that of the CRT-CA. In short, depending upon relative discriminatory impact, certification officials who value knowledge test scores could use either or both the NTE or more specialized tests developed to measure knowledge deemed relevant to local needs. However, considering the results of this and previous studies, one must conclude that knowledge tests of the sort used now and in the past provide little information as to a person's effectiveness as a teacher. Thus it seems that unless new tests cover other types of knowledge, development of more knowledge tests for the purpose of evaluating teachers may be a superfluous activity.

Implications

The results of this study along with information discovered while conducting the study hold implications for program development and for future research. Both types of implications merit discussion.

Implications for Program Development

The implications for program development involve suggestions for content emphasis, for evaluation of TC performance, and for selection of instruments. This study does not yield information as to the cause of the negative relationship between Tested Knowledge and Professional Ratings, but research discussed by Flanders and Simon (1969) may provide some insight. Sponsored by the government of Finland, Koskenniemi conducted a longitudinal study in which he discovered traits common to teachers with poor inservice records. These

teachers lacked understanding of children's attitudes and thinking and lacked the ability to structure effectively instructional situations (Flanders & Simon, p. 1432). Possibly, both of these shortcomings indicate a lack of knowledge, but a knowledge of content not usually covered by knowledge tests such as the NTE, and even by specialized knowledge tests like the CRT-CA. Perhaps in campus and field-based courses more emphasis should be placed on knowledge of findings from research in teacher effectiveness and research in teaching of English. Hopefully, increased knowledge of research on teaching composition, language, and literature would help TC's structure their instruction more effectively. Whether or not this shortcoming stems from lack of this type of knowledge, increased help and practice in planning instructional strategies should help improve a TC's classroom performance.

In terms of evaluating TC's, this study provides no conclusive evidence as to the effectiveness of the predictors or criterion measures used by the Lanuage Education Department. While none of the measures predict successful Professional Ratings, the program must continue to measure growth and to evaluate TC's and the program. Moore (1976) concludes that since no single criterion measure covers all of the program objectives, no one instrument can serve as criterion (p. 94). This conclusion warrants no challenge on the basis of findings from this study; further, Moore's suggestion to utilize several different types of criterion measures coincides with the opinions of experts on teacher effectiveness research and applies to any teacher preparation program.

Ideally, a program should include pupil outcomes as criteria for evaluating TC's. If adequately reliable and valid measures of pupil knowledge and attitudes could be obtained, and if research and/or statistical designs could be employed to eliminate the influence of other variables besides those controlled by the teacher, and if the use of such measures were economically feasible, a program

would be remiss in not employing pupil achievement as a criterion. Since none of these contingencies has been successfully resolved by the hundreds of other teacher effectiveness studies, this program and others like it can hardly be chided for the omission. However, pupil behavior could be examined to a greater extent with little trouble. It would take little effort to add some indications of pupil behavior to existing observation and rating forms and to train supervisory team members to record these behaviors in more detail. However, any more thorough treatment of pupil behavior would require another instrument. Further, students' feelings about their classes should continue to be consulted despite the failure of this study and others to uncover relationships between Student Ratings and other variables. Again, all types of criterion should be used where feasible.

The implications for the selection of instruments merit discussion but seem rather impractical. The need for reliable instruments to measure process variables has been illustrated over and over again, and ideally any program would use highly reliable and valid instruments to measure presage, process, and product variables. Unfortunately, videotaping and other mechanical means of recording classroom events are as expensive as they are accurate. Other methods of collecting data may be less expensive yet still fairly accurate, but impractical for use by supervisors who must observe many TC's each week.

Therefore, live observation by the supervisors is the only viable method. The instrument used by these observers to encode classroom events and rate the TC must serve two functions. If the instrument involves an elaborate coding of only a few selected behaviors, it will be of little use to supervisor or TC in subsequent discussions. If the instrument contains only ratings, it will provide few concrete examples of the teaching process which was rated. A combination observation/rating form, like the TC Observation Form used in this program, suffers

from the faults of both types of instruments. Nevertheless, the utility of this type of form and its summary report for the TC's and supervisors has been demonstrated in practice, and while this sort of instrument may not be ideal for use in a research design, it can justifiably be retained for use in teacher preparation programs and in research related to these programs.

The same assessment holds for all the instruments used in this study which ideally could be replaced by instruments more suitable for research on teacher effectiveness. However, the goals of this program focus on teacher preparation, and while the accumulation of data for research on teaching is a valuable byproduct, the choice of instruments and procedures depends on the impact of these choices on the TC's and their student teaching experience rather than on the soundness of research designs. Nevertheless, valuable information and data can be obtained from such instruments and such programs; this information and data should not be neglected because of less than perfect research conditions. Few studies of teacher effectiveness have enjoyed the luxury of financial and administrative independence that allows the researcher to design a clean experimental study. The goal of teacher effectiveness research is to improve teaching, which after all takes place in real classrooms, in real schools, with real students from the real world, and worthwhile research findings will have to generalize to these existing situations.

Implications for Future Research

Future research may benefit from some specific suggestions and from some general ones generated by this study and prior research. Perhaps further investigation of the relationship of presage variables, such as teacher knowledge and attitudes, to process and product variables would accomplish more if instruments were constructed in the way Dunkin and Biddle suggest, by basing these

instruments on observations of classroom behaviors and working backwards to develop the instruments (1974, p. 442). Stronger relationships might be discovered if measures of knowledge covered knowledge of research on teaching effectiveness and research on teaching specific subject matter and processes, for example language, literature, composition, etc. Surely, for example, knowledge of research on composition can aid the teacher in becoming a more effective teacher of writing.

In general, teacher effectiveness research would benefit from the suggestions for new directions offered by such experts as Gage, Soar, Harutoonian, Joyce and Medley, who share Dunkin and Biddle's concern over the failure of instruments and models to derive from realistic conceptualizations of teaching, a failure which has plagued CBTE advocates in their search for a unified set of validated competencies. Future research may prove more productive if it focuses on matching conditions, procedures, purposes, and students to discover elements of effective teaching rather than continuing to search for the one set of traits and competencies which characterize the effective teacher. The agreement among these experts as to the need for and nature of this new direction focuses attention again on research on teaching of specific subject matter and processes. Certainly the kind of matching models research approach suggested above would require the researcher investigating English teacher effectiveness to look more closely at the vast store of knowledge amassed in recent years by research in composition, language, literature, and reading.

Summary

Research growing out of teacher preparation programs should continue to utilize existing approaches and should explore new approaches; researchers should not abandon this field despite difficulties in collecting data, obtaining perfect instruments, devising sound research designs, and discovering significant

relationships among presage, process, and product variables. Neither should teacher educators despair over the failure of research to prove empirically the validity of their preparation programs. Teachers must be trained to teach as best they can, and while empirical proof of that teaching ability should be sought, schools can hardly postpone their tasks until that proof is found. If follow up studies show that former TC's are performing well enough in classrooms to satisfy parents, administrators, and students, then conclusive empirical proof will simply add to the already justified satisfaction of teacher educators.

References

Angoff, W.H. Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Broudy, H.S. A critique of performance-based teacher education. Washington, D.C.: AACTE, 1972.

Claudy, J.G. An empirical investigation of small sample multiple regression and cross validation. Unpublished doctoral dissertation, University of Tennessee, 1969.

Cook, W.W., Leeds, C.H., & Callis, R. Minnesota teacher attitude inventory. New York: The Psychological Corporation, 1951.

Dunkin, M.J., & Biddle, B.J. The study of teaching. New York: Holt, Rinehart, & Winston, 1974.

Flanders, N.A. & Simon, A. Teacher effectiveness. In R.L. Ebel (Ed.), Encyclopedia of educational research (4th ed.). New York: MacMillan, 1969.

Gage, N.L. Evaluating ways to help teachers behave desirably. In W.R. Houston (Ed.), Competency assessment, research, and evaluation: A report of a national conference. Houston, Texas, March, 1974.

Georgia Department of Education. State department of education progress report on performance-based certification in Georgia. April-May 1977.

Getzels, J.W., & Jackson, P.W. The teacher's personality and characteristics. In N.L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

Harootunian, B. Research on teaching effectiveness. In M. Mohan & R. E. Hull (Eds.), Teaching effectiveness. Englewood Cliffs, N. J.: Educational Technology Publications, 1975.

Heath, R. W., & Nielson, M. A. The research bases for performance-based teacher education. Review of Educational Research, Fall 1974, 44 (4), 463-484.

Joyce, B. R. Listening to different drummers: Evaluating alternative instructional models. In W. R. Houston (Ed.), Competency assessment, research, and evaluation: A report of a national conference. Houston, Texas, March 1974.

McDaniel, T. R. The NTE and teacher certification. Phi Delta Kappan, November 1977, 59 (3), 186-188.

Medley, D. M. Closing the gap between research in teacher effectiveness and the teacher education curriculum. Journal of Research and Development in Education, Fall 1973, 7 (1), 39-46.

Medley, D. M. The research context and the goals of teacher education. In M. Mohan & R. E. Hull (Eds.), Teaching effectiveness. Englewood Cliffs, N. J.: Educational Technology Publications, 1975.

Medley, D. M., Soar, R. S., & Soar, R. Assessment and research on teacher education: Focus on PBTE (PBTE Series No. 17). Washington, D. C.: AACTE, June 1973.

McNeil, J. D. & Popham, W. J. The assessment of teacher competence. In R. M. W. Travers (Ed.), Second handbook of research on teaching. Chicago: Rand McNally, 1973.

Mitzel, H. E. Teacher effectiveness. In C. W. Harris (Ed.), Encyclopedia of educational research (3rd ed.). New York: MacMillan, 1960.

Moore, C. A. Evaluation of the pre-service secondary English education program at the University of Georgia. Unpublished doctoral dissertation, University of Georgia, 1976.

Nassif, P. M., & Dewitt, G. W. The Georgia teaching field criterion-referenced test project: Final report. Amherst, Mass.: National Evaluation Systems, June 1977.

Popham, W. J. Designing teacher evaluation systems. Los Angeles: Instructional Objectives Exchange, 1971.

Quirk, T. J., Witten, B. J., & Weinberg, S. F. Review of studies of the concurrent and predictive validity of the National Teacher Examinations. Review of Educational Research, Winter 1973, 43 (1), 89-113.

Rosenshine, B., & Furst, N. Research on teacher performance criteria. In B. O. Smith (Ed.), Research in teacher education: A symposium. Englewood Cliffs, N. J.: Prentice-Hall, 1971, 37-72.

Schutz, W. C. FIRO-B. Palo Alto, Calif.: Consulting Psychologists Press, 1957.

Scott, O. & Veal, L. R. The inventory of student perception of instruction. Unpublished, University of Georgia, 1970.

Secondary English Committee, Department of Language Education. The University of Georgia teacher education program in English. English Education, Summer 1976, 7 (4), 218-235.

Soar, R. S. Multivariate statistical procedures in predicting teacher-pupil classroom behavior (Cooperative Research Project No. 1170). Columbia, S.C.: University of South Carolina, 1962.

Study Commission on Undergraduate Education and the Education of Teachers. Teacher education in the United States. Lincoln, Nebraska: University of Nebraska Press, 1976.

Tractenberg, P. L. Testing the Teacher. New York: Agathon Press, 1973.

Uhl, N.,& Eisenberg, T. Predicting shrinkage in the multiple correlation coefficient. Educational and Psychological Measurement, Summer 1970, 30 (2), 487-489.

Veal, L. R., & Biesbrock, E. The development of an experimental, standardized instrument for measuring composition ability in young children (Technical Report No. 9). Athens, Ga.: University of Georgia, Research and Development Center in Educational Stimulation, October 1969.

Veal, L. R., Pikaart, L., & Berryman, C. Improving student teaching through objectivity and research. Bulletin of the bureau of school services, March 1967, 39 (3).

Villeme, M. G. The decline of competency-based teacher certification. Phi Delta Kappan, January 1977, 58 (5), 428-429.

Vlaanderen, R. B. State review: Focus on bargaining and certification. Legislative Briefs, February 1978, 4 (2), 6-7.

Ward, B., Morine, G., & Berliner, D. C. Assessing teacher competence. In W. R. Houston (Ed.), Competency assessment, research, and evaluation: A report of a national conference. Houston, Texas, March 1974.